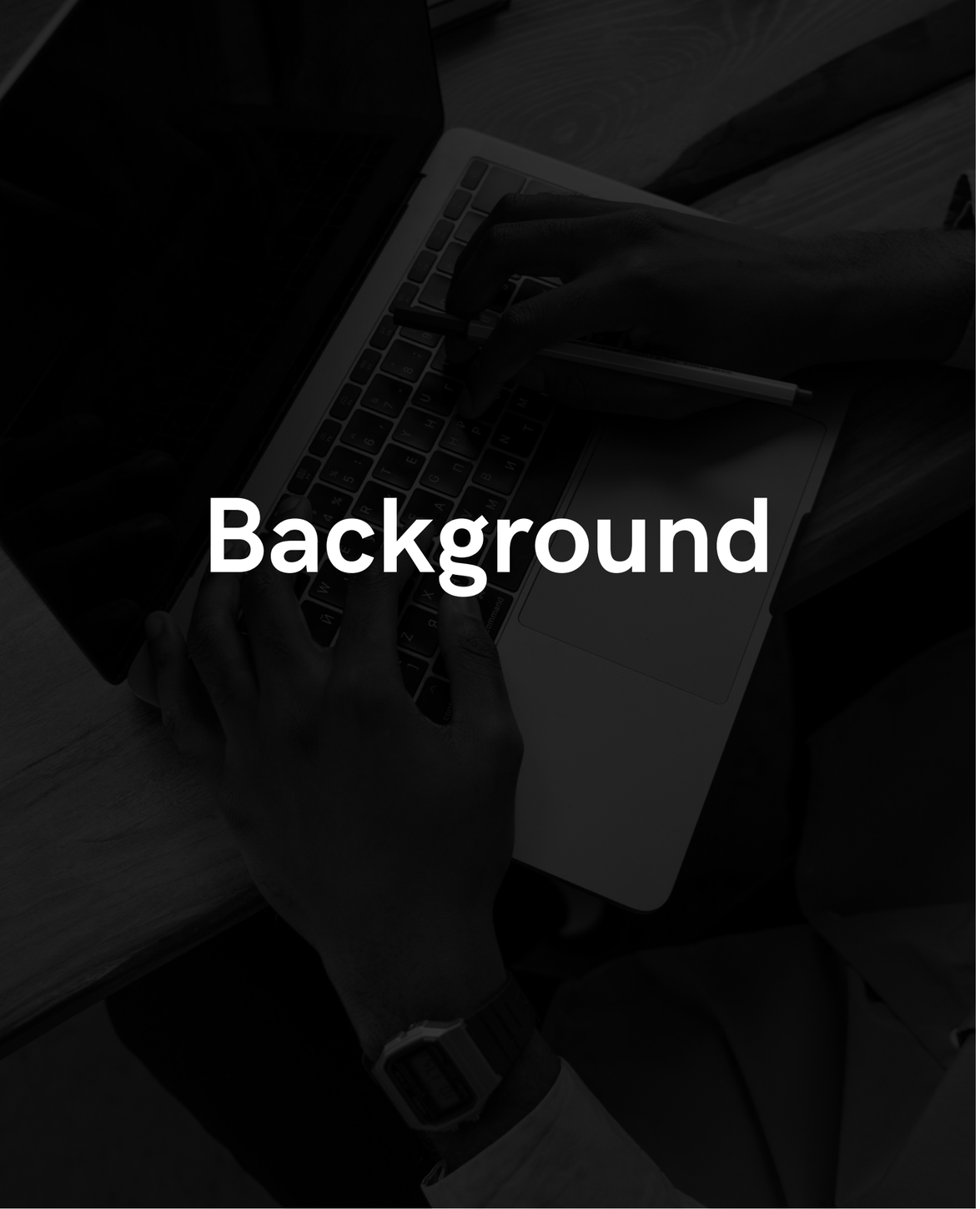

# PREDICTING HIGH-COST PATIENTS IN HEALTHCARE USING PREDICTIVE ANALYTICS

-SAMBHAV JAIN



# Background

## Health Expenditure

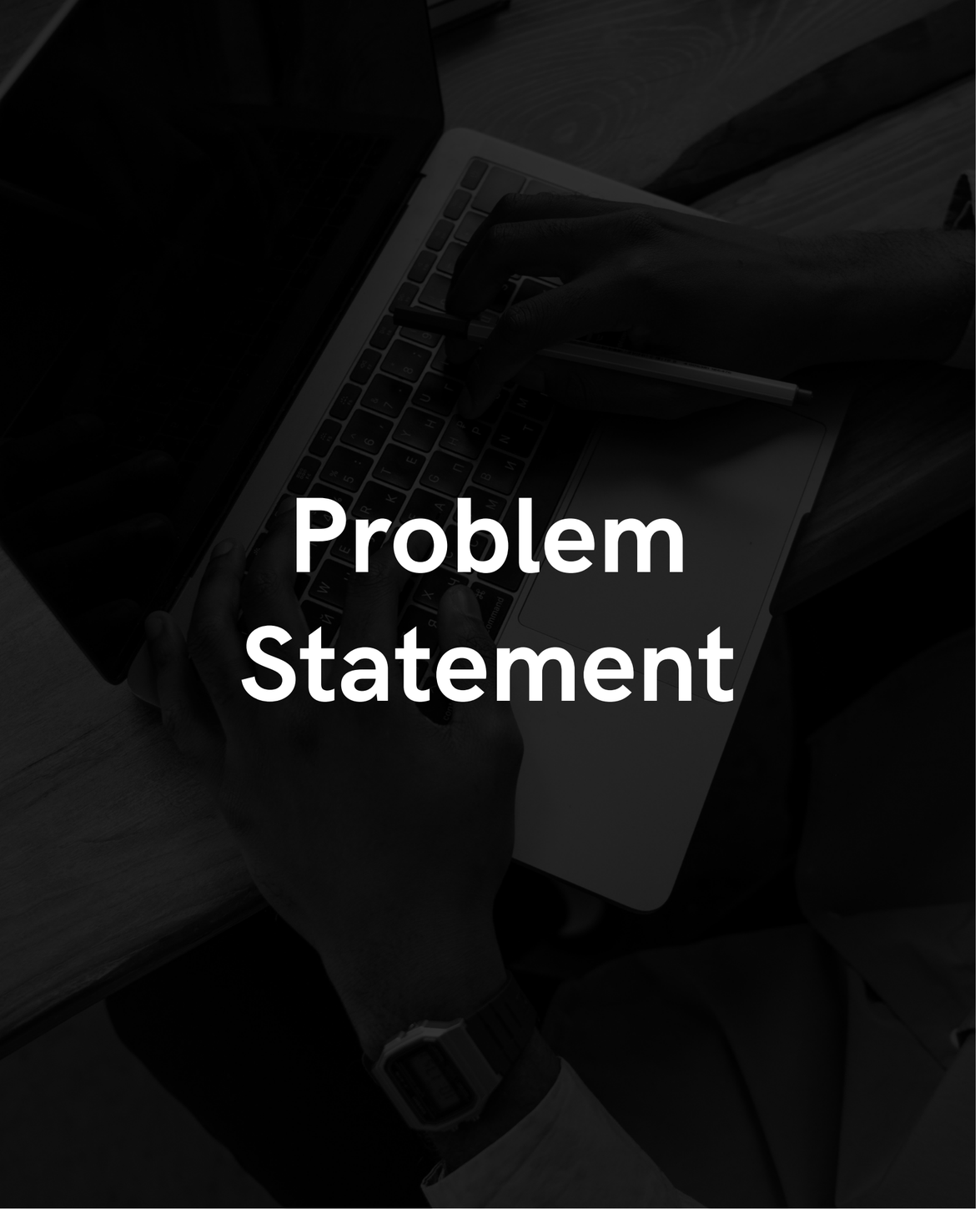
- Every nation facing the issue of rise in health expenditure
- A small proportion of patients bear large chunk of expenditure. These are called “High-cost” patients
- Identifying high-cost patients is a top priority and predicitive analytics can help in this work

## Limitations of Traditional Models

- Traditional Methods use simpler methods like regression
- They miss out on non-linear and complex patterns
- Traditional models often struggle with diverse dataset

## Role of ML

- ML models can handle non-linearity and interactions between datapoints
- Helps in improving the prediction accuracy of the model
- Faces challenges like data inconsistency, poor interpretability, and limited focus on patient classification



# Problem Statement

## Limitations of Existing Models

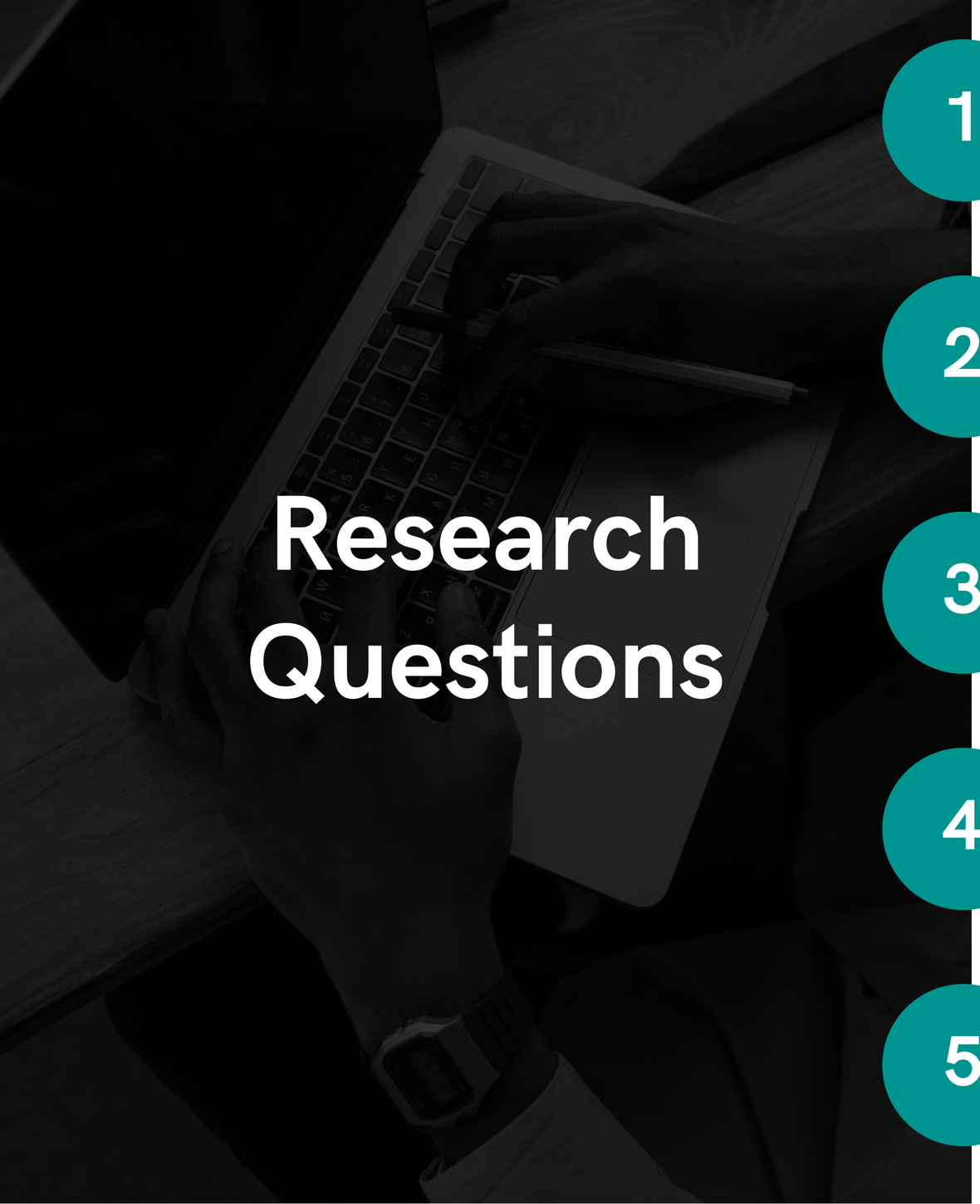
- Traditional regression models can't capture non-linear patterns
- Machine learning models (e.g., XGBoost, SVM) improve accuracy
- But lack interpretability for real-world decision-making

## Data and Classification Challenges

- High-cost patients form a minority showcases class imbalance
- Models often biased toward low-cost majority
- Leads to poor identification of true high-risk individuals

## Gaps in Segmentation and Explainability

- Clustering methods like K-Means, hierarchical underutilized
- Cost stratification not integrated into predictive frameworks
- Few models use SHAP to highlight key cost drivers



# Research Questions

1. How can ML models predict high-cost patients using demographic, clinical, and socio-economic data?
2. Which features most influence healthcare costs, and how can SHAP improve feature selection and interpretability?
3. How can clustering help segment patients into cost tiers for better resource allocation?
4. How do classification models compare with regression models in predicting costs accurately and interpretably?
5. How do data handling strategies—like excluding target-derived variables or managing class imbalance—impact model fairness and performance?

# Aims and Objectives

- Assess existing statistical methods for predicting high-cost patients and identify their shortcomings
- Use machine learning models—especially XGBoost—to enhance prediction accuracy
- Leverage SHAP for feature selection to uncover the most impactful variables affecting healthcare costs
- Combine ML with explainability tools to ensure transparency and clinical relevance
- Validate models using cross-validation and performance metrics like accuracy, precision, recall, AUC, and RMSE

# Significance and Scope

## Significance of the Study

- Enables early identification of high-cost patients for timely intervention
- Helps insurers design better premium and claim models
- Aids policymakers in optimizing healthcare resource allocation
- Enhances trust and usability through interpretable ML models (e.g., SHAP)

## Scope of the Study

- Focuses on ML-based prediction of high-cost patients using structured inpatient data
- Covers data preprocessing, feature selection, model building, and evaluation
- Excludes real-time systems, ethical policy design, and clinical treatment recommendations
- Targets practical, explainable models for hospitals, insurers, and policymakers

# Literature Review

## High-Cost Patients and Cost Concentration

- A small fraction of patients (~5%) incur ~50% of healthcare costs
- These patients often have chronic diseases and complex social needs
- Early identification enables targeted, proactive care interventions

## Limitations of Traditional Methods

- Rely on static models (e.g., regression, risk scores) assuming linear relationships
- Cannot adapt to changing conditions or diverse data inputs
- Often suffer from label leakage and fail to distinguish avoidable vs. unavoidable costs

## Rise of ML in Cost Prediction

- ML handles non-linear, high-dimensional data better than traditional methods
- Techniques like XGBoost and deep learning boost accuracy and scalability
- Challenges remain: model interpretability, overfitting, and clinical adoption barriers

# Literature Review

## Real-World Applications of ML

- Claims-Based Models uses structured data but lack social context
- EHR-Based Models capture clinical nuance but suffer from fragmentation
- Hybrid Models improve accuracy by combining claims and clinical insights or focusing on specific conditions

## Importance of Feature Engineering

- Key Features like Demographics, diagnosis codes, utilisation metrics are core predictors
- Encoding techniques (e.g., one-hot, embeddings) tackle sparsity
- Non-clinical factors (e.g., income, housing) enhance fairness and relevance

## Role of Explainability in Healthcare ML

- Stakeholders require transparent, auditable models to build trust
- Tools like SHAP & LIME: Help explain model predictions at global and individual levels
- Identify over-reliance on proxies and enable ethical use of ML in cost prediction

# Literature Review

## Comparative Performance

- Ensemble models (e.g., XGBoost) outperform traditional regression in accuracy and flexibility
- Deep learning handles temporal data but faces adoption barriers due to opacity
- Institutional case studies highlight context-specific challenges

## Advances in Modern ML Techniques

- Temporal models (RNNs, LSTM) improve prediction of chronic vs. episodic cost patterns
- Reinforcement and multi-task learning enable cost-sensitive, decision-focused care optimization
- Drift detection systems ensure model adaptability and reliability over time

## Policy, Practical, and Ethical Imperatives

- Fairness and privacy audits are essential to detect bias and protect sensitive data
- Clinical integration requires explainability, provider alignment, and distinction between avoidable/unavoidable costs
- Deployment challenges include latency, accountability, governance, and regulatory alignment

# Gaps in existing Literature

Many models depend on prior expenditure data, which risks label leakage and limits real-time applicability in predictive systems.

Most studies focus on regression-based predictions, overlooking classification into cost tiers which is more actionable for healthcare triage.

While tools like SHAP exist, their integration into high-cost patient models is sparse, making outputs less interpretable for stakeholders.

Models are often trained and tested within a single dataset or health system, restricting their validity across populations and care environments.

Despite high accuracy, few models are deployed in live healthcare systems due to issues with interpretability, clinical integration, and actionability.

# Dataset Selection & Handling

## Dataset Selection

- Hospital Inpatient Discharges data by SPARCS selected for research
- With over 2.1 million rows and 33 columns, it offers robust foundation
- Covers granular details related to demographic, clinical, diagnosis and procedure of the patients
- Ensures patient privacy through de-identification, thereby adhering to ethical standards for data-driven research.

## Data Handling

- Cleaning of data to eliminate oddities
- Handling missing values either through imputation or removing them altogether
- Detection of outliers and transformation of variables
- Feature Engineering and Categorical encoding
- Univariate and Bivariate analysis of features to understand their nature

# Model Selection and Evaluation

## Model Selection

- Ridge and XGBoost regression were used to predict total costs, but dropped post-EDA due to high number of categorical variables in the dataset

- XGBoost classifier grouped patients into cost tiers, offering better interpretability and practical use in care planning

## Model Evaluation and Interpretation

- Using Accuracy, precision, F1-score, Recall, AUC-ROC to evaluate classification modelling
- K-fold cross-valuation to ensure no overfitting of the model
- SHAP analysis to understand contribution of features in the prediction

# Model Results of Base Model

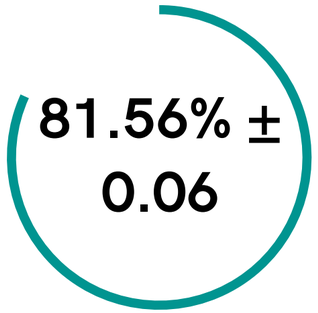

Accuracy


Precision


Recall


F1-score


AUC-ROC




Cross-Val Accuracy

Evaluation metrics of the classification model showcases that the strong predicitve performance of the model

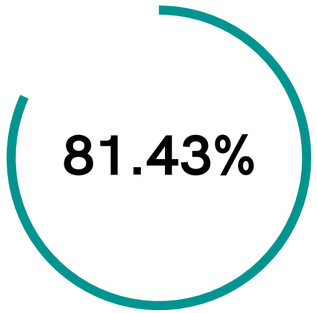
SHAP analysis reveals that total estimated cost has a very siginifcant influence on predicting power of model.

Other than total estimated costs,  variables like length of stay, cost per day, Hospital Service Area and APR Medical Surgical Description are very signifcant

Need to develop an alternate model without total costs to understand underlying patterns and assess model’s robustness without a variable having strong relation to target variable

# Model Results of Alternate Model


Accuracy

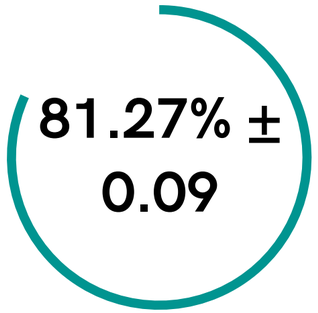



Precision


Recall


F1-score


AUC-ROC




Cross-Val Accuracy

Evaluation metrics of the Altenate classification model are little lower than base model but still showcases that the strong predicitve performance of the model

SHAP analysis reveals that length of stay, cost per day, Hospital Service Area, APR Medical Surgical Description, Emergency Department Indicator, and Severity of Illness influence the most

As alternate model avoids information leakage, Offers superior explainability, aligns with real-world clinical workflows, and enables earlier risk detection, it is  recommended as the more robust and deployable solution

# Discussion and Conclusion

ML models can classify high-cost patients using structured inpatient data, even without actual cost values

The study highlights the trade-off between model accuracy and interpretability in real-world applications

The interpretable model offers ethical, point-of-care deployability despite slightly lower precision

Cost-tier classification supports planning, forecasting, and risk management in healthcare operations

# Knowledge Contribution

## Healthcare Systems and Hospitals

- Showcased that hospitals can classify patients using admission-time features
- ready-to-integrate model structure that aligning with operational workflows

## Policymakers and Authorities

- Contributed ethical framework for early risk prediction
- Prepares case for using interpretable models in cases where cost data may not be available

## Data Science Teams

- Provided a scalable, modular framework on structured hospital data
- Use of proxy variables for unavailable or sensitive variables

## Academia and Educators

- Introduced new perspective of treating cost prediction as a classification problem
- Validated models without target-adjacent features and displayed use of SHAP

# Future Recommendations

## Healthcare Sector

- Intergration with EHRs can trigger alert for high-risk patients
- Pair model outputs with intervention protocols for early discharges

## Insurers

- Use cost classification to priroitise approvals for high-risk cases
- Use model insights for preventive outreach or bundled payment strategies

## Data Science Teams

- Improve the accuracy of the model by incoproating temporal data
- Test for any performance disparities and correct for systematic biases

## Academics

- Integrating the model in healthcare analytics and applied machine learning courses
- Ecourage projects connecting healthcare, economics and ML

## Policymaking

- Testing the model in public healthcare to check for impact at larger scale
- Support guidelines for responsible use of AI in healthcare